Washington State Archives, Digital Archives

The evolution of ingestion

Presented by:

Daniel Waterbly
dan.waterbly@sos.wa.gov

September 2010, Cheney WA
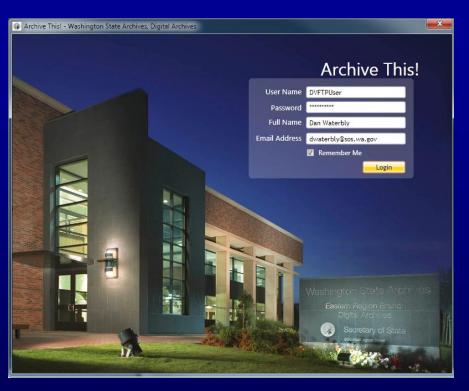
Washington State Archives   *Documenting Democracy*

Archive This!

- Supports many data formats
- Validates raw data on partner's machine
- Structures incoming data into common file structure
- Transfer data to hard drive or allows data to be uploaded over FTPS
- Fingerprints all incoming data before we receive it at the Digital Archives
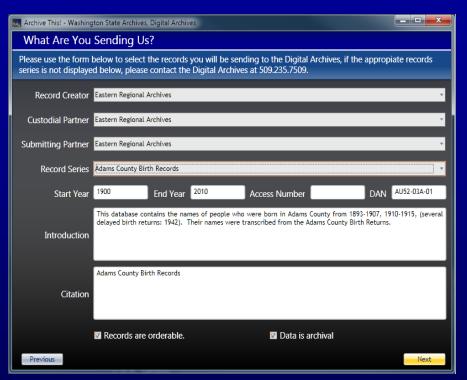- Gateway to Ingestion System

Archive This

Archive This

Archive This

Washington State Archives   *Documenting Democracy*

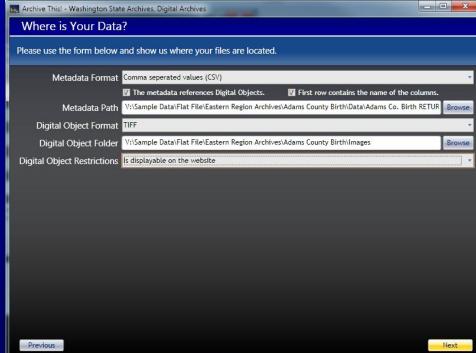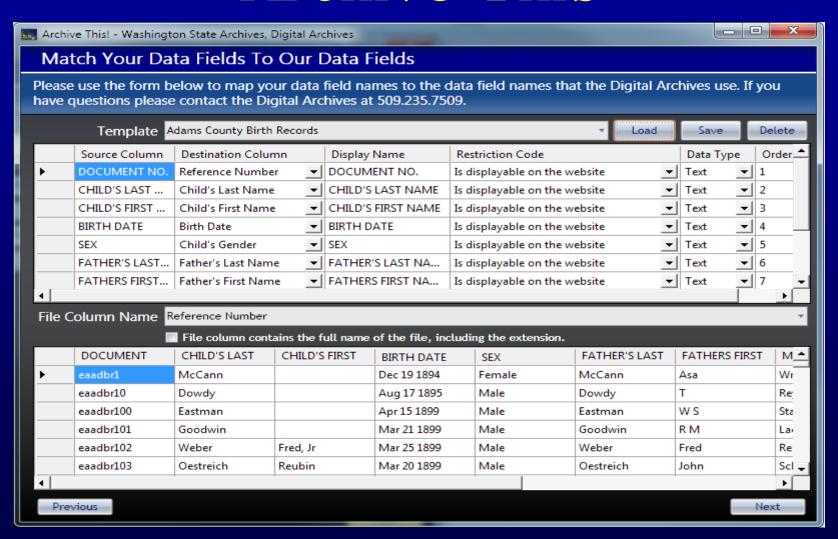# Archive This

The Evolution

- Originally WADA Data Transfer
  - Had following Issues
    - No data validation
    - No update capability
    - No upload/export support
    - Column mapping was hard

How is it implemented now?

- Written in C#

- Uses WPF for user interface

- Uses WCF for interaction with the DA database

- File transfer takes place using FTPS (block mode)

Highly Desirables

- Automate application on client's server
- Add a new record series without a deployment and code change

Auto Todd

- Monitors incoming data

- Virus scans new incoming data

- Verifies that new incoming data is the same on our server as it was on the client's server

- Processes new incoming data

- Sends alerts based on process state

- Backs up data to remote storage area

- Tracks all incoming data and all of its states as it travels through the ingestion system

The Evolution

- First iteration used SSIS

- Everything was manual

- Data validation and file matching was done on the Digital Archives' end

- Required much human interaction

How is it implemented now?

- Consists of several Windows services written with custom C# code

- Each service hosts a workflow which outlines how the specific service will work

Main Processing

- Ingests individual records in the archives
- Distributed on sixteen quad core servers and two eight processor servers
- Converts files to presentation format
- Extracts text from audio
- Each record either goes in completely or not at all (atomic)

Contacts

Dan Waterbly

dan.waterbly@sos.wa.gov